PSTL's BN-STT system

Panasonic
PSTL

Patrick Nguyen
Jean-Claude Junqua

Panasonic Speech Technology Laboratory
(PSTL)

Plan

- System description
- Summary of improvements
- Large corpus experiment

System description

- Same as last year: 2-pass word-internal GD trigram Viterbi (EWAVES)
- Improvements made on the *models*

Summary of improvements

- Official RT02 results
  - 10xRT: 20.1% WER
  - 1xRT: 23.7% WER
- RT03S system, RT02 set: RT03evl
  - 10xRT: 16.1% WER 15.2% WER
  - 1xRT: 19.8% WER 20% WER

- 20% WER improvement or 10x in speed

Strategy

- Spend 60% on system development
- Spend 40% on “new features”

Improvements

• Last year's system	20.1% WER
• Tuning & retraining	19.6% WER
• MLLU features	19.0% WER
• TDT/MMI	17.5% WER
• MLLU adapt	16.8% WER
• LM	16.1% WER
• Reseg TDT (post-eval)	15.3% WER

Improvements

• Last year's system	20.1% WER
• Tuning & retraining	19.6% WER
• MLLU features	19.0% WER
• TDT/MMI	17.5% WER
• MLLU adapt	16.8% WER
• LM	16.1% WER
• Reseg TDT (post-eval)	15.2% WER

=> 2.4% absolute from MMI on large corpus

Large Corpus

- Recent CoreTex research
- 10k hours corpus collection has begun

- Statistical learners are slow
- Much time spent in smoothing algorithms
- Let the machines do the thinking
- Isolet syndrome: low portability
- Over-training in general

TDT Collection

- About 1400h of data, 38M words
- TDT2: 550h, 20M words
- TDT3: 400h, 9M words (delete Dec 1998)
- TDT4: 350h, 9M words

versus

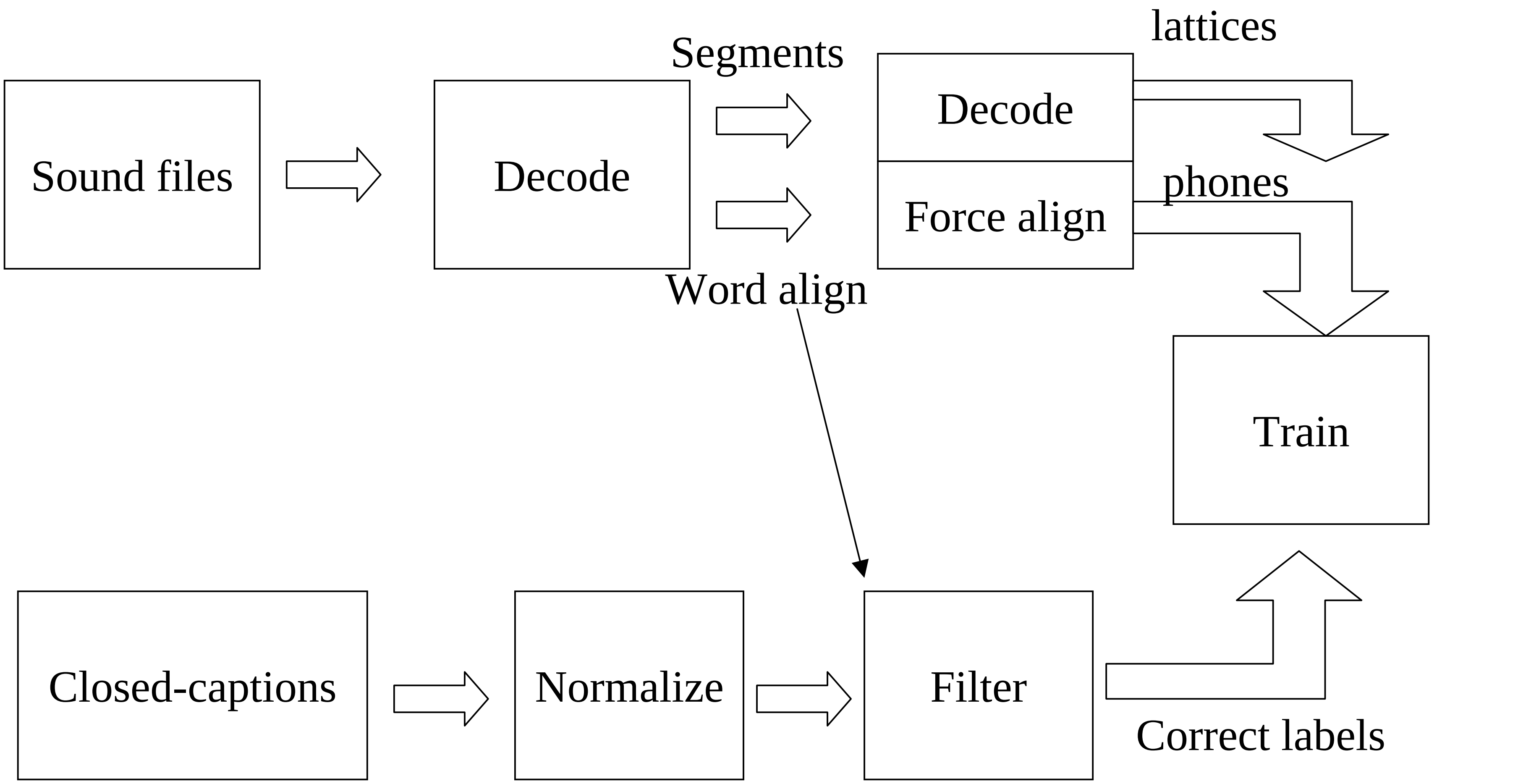
- Hub4: 200h, 1.2M
- One order of magnitude

Lightly supervised training [LIMSI]

- TDT2: 550h
- Match the baseline, ignore Hub4train
- No discriminative training
- Filtering is different
- Iterate many times

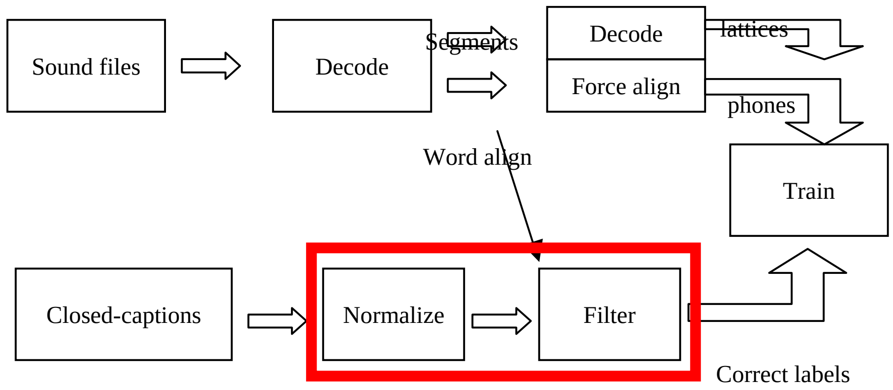
- All tokens are trainable: breath, cough, etc.

TDT processing



Text processing

- From captions to ASR transcripts
- Reverse MDE task
- Our standard LM normalizer



Erosion filter

- Cross-word contexts (ripple effect)
  - If no match, then probably wrong context
- Time alignment of wrong words
  - Corrupt the alignment of neighboring words
- DP match is too “nice”
  - E.g.: the **the** e. e. **a** a.

Good Monday evening zero there are signs that the

Biased training

- Amount of training data depends on language model probability
- => apply the LM twice?
- Amount depends on quality of speech (recognition results)
- Depends on prior probability in general (Male/Female)
- We ignore these issues

Error-proof training

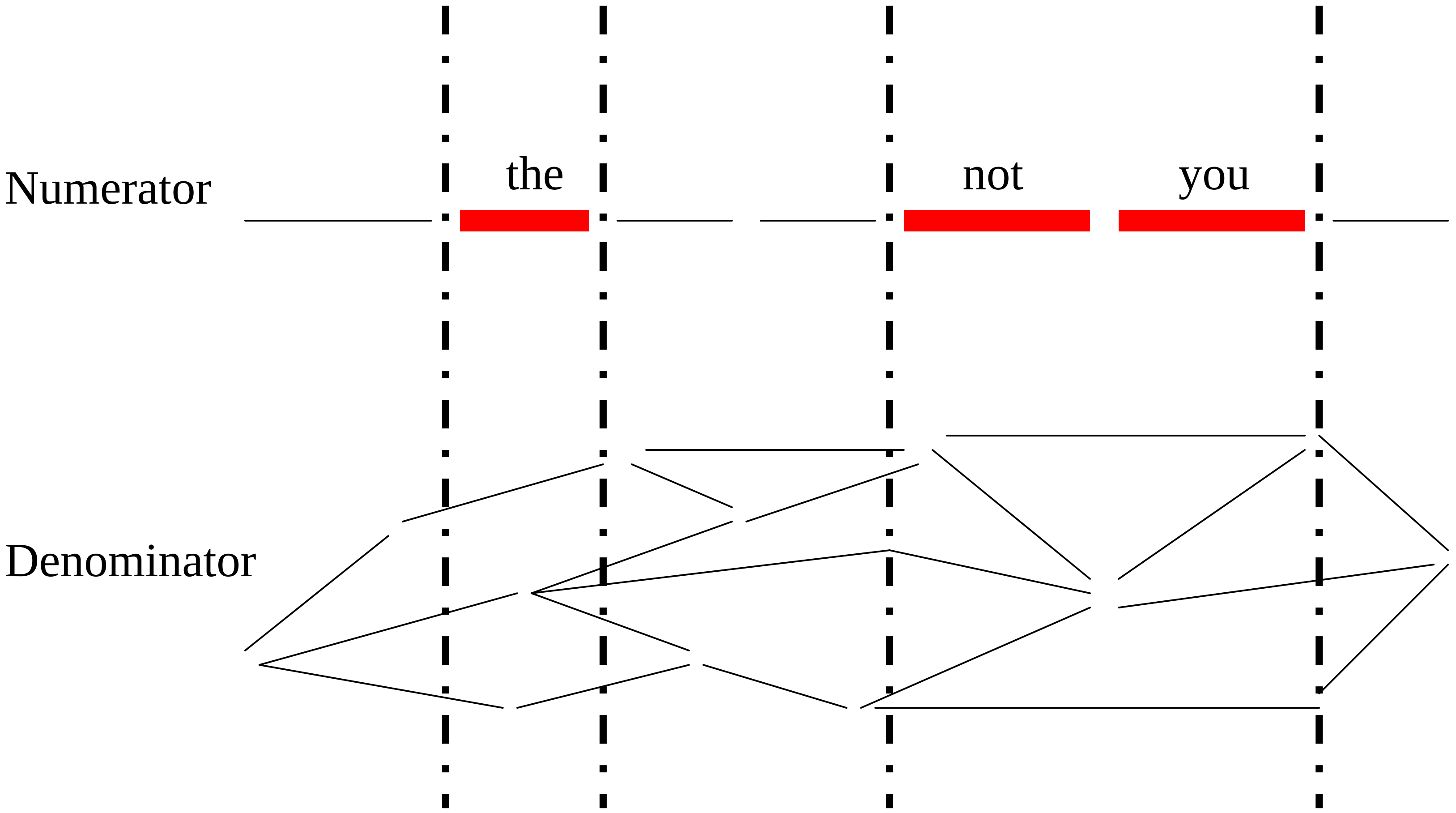
- Manual processing is not practical
- Major difficulty in large amounts: outliers
- Murphy’s law (NFS, max inodes, …)
- Crash, fix, and retry is not practical
- Simple rule: DISCARD
- Error-proof training

Incorporation of new data

- TDT4 arrives in PSTL on April 4, 2003.
- Decodings: 5h (1-best), 7h (lattice gen)
- Start with 1-iteration MMI models
  - 15h / iteration
- 30h + 12h + crash + processing

- Integrated 350h of data in one week-end with error-proof training
- Many thanks to Stephanie Strassel and publishing group at LDC!

MMI



Integration of the denominator is non-trivial, but does not matter

Scalability: orders of magnitude

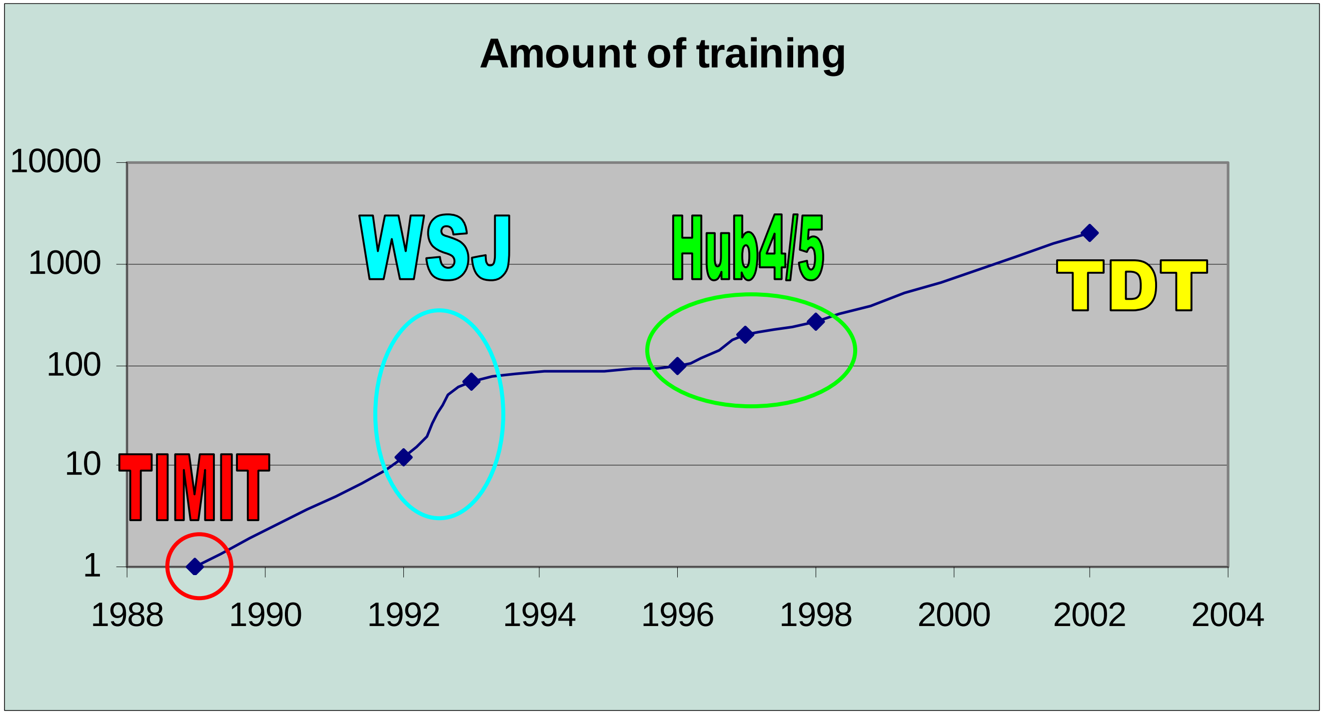
- Total training: 1600h
- Utterances: 500k
- Labels: 6 GB
- Lattices: 98 GB
- Current cluster: 88 CPU

- Decode (lattice gen) 50h audio / h cluster
- MMI:

		Raw	Filtered
– Females:	6h / it	409h	272h
– Males:	9h / it	1028h	687h

Towards 10’000h corpus?

- 360h / mo (LDC)
- Two years (>24mo) to complete (2005)



LinLog Slowdown Rule

- Hypothesis: 6x data (1600h=>10000h)
- Training is linear: 6x
- More Gaussians: log(6x)

- But:
  - Moore Law (exp) vs linear LDC collection (lin)

    2yr: CPU x 4, data x 6
  - Algorithmic improvements (linear?)
- **Simplify**, rather than complicate, training (e.g. absolute discounting vs Turing-Good)

TDT: Conclusion

- Scaled up standard training techniques
- Successful particularly with data savvy MMI and Gender Dependent
- XW pentaphones, SAT not considered yet
- silence, word fragments not considered
- Smoothing tuning disappears
- This is merely the beginning…

Fiscus-Moore Effect

- Fiscus: variability is good
  - [Schwenk & Gauvain 2000: Improving ROVER]
- Moore: 2 x 10xRT now = 10xRT next year

- Can guarantee 10% relative improvement for two consecutive years
- => in 2005, 8.6% WER @ 9xRT w/o much work if team up or share resources

Fiscus-Moore: results

- ½ of RT03S (spkr set), ROVER=0xRT
- BBN: 10.8%, LIMSI: 10.8%, SRI: 13.4%, CU: 10.4%, CU-1x: 14.2%
- RT04:
  - BBN+LIMSI: 9.7% (17.5xRT)
  - BBN+LIMSI+CU-1x: 9.3% (18.4xRT)
- RT05:
  - BBN+LIMSI+SRI+CU: 8.6% (36.2xRT)

[Thanks to Phil Woodland and Jon Fiscus for providing CTMs]

Conclusion

- 25% improvement since last year
- Large corpus experiment
- Other improvements from MLLU, LM
- Word internal decoder
- More contributions

END

Any questions?

MLLU

- Maximum-likelihood Lower-Upper transformation
- Presented at ICSLP02

- Closed-form solutions for linear feature transformation
- Problem similar to matrix inversion (logdet)
- Better control than Laplace expansion